УДК 004.934

Н.Б. Васильєва, М.М. Сажок
Міжнародний науково-навчальний центр інформаційних технологій і систем,
м. Київ, Україна
ninel@uasoiro.org.ua, mykola@uasoiro.org.ua

Моделювання багаторівневого поскладового розпізнавання мовленнєвого сигналу

В статті проводиться поширення багаторівневої багатозначної моделі автоматичного розпізнавання злитого мовлення на випадок поскладового розпізнавання. Розглядаються два рівні з трьох. На першому рівні проводиться розпізнавання в умовах поскладової граматики, на другому рівні проводиться оброблення (постпроцесинг) вихідних даних першого рівня з метою отримання відповідних послідовностей слів. В описаній моделі постпроцесингу беруться до уваги отримані оцінки акустичних складових мовленнєвого сигналу, а послідовність і фонетичні особливості разом з лексиконом. Аналізуються шляхи вибору множини одиниць на складовому рівні мовленнєвих образів. Описується багатодикторний мовленнєвий корпус і лексикон, що використані в експериментальному дослідженні. Обговорюються результати експериментів, проблеми та майбутні дослідження.

Вступ

У системі багаторівневого багатозначного розуміння мовленнєвого сигналу, описаній у [1], злите мовлення спершу розпізнається як послідовність фонем, а потім ця послідовність фонем перетворюється на послідовність слів та проводиться смислова інтерпретація.

Незважаючи на те, що найкращий метод розуміння мовленнєвого сигналу полягає в його одночасному розпізнанні та смисловій інтерпретації, конструювання такої багаторівневої системи є можливістю розподілити науково-дослідну роботу між експертами в акустиці, фонетиці, лінгвістиці та інформатиці. Очевидно, що багаторівнева структура смислової інтерпретації мовлення є найбільш продуктивною при створенні систем диктування та систем усного діалогу для ряду надзвичайно флективних мов з відносно вільним порядком слідування слів, до яких відносяться і слов’янські мови.

Слід зауважити, що результат дії кожного з рівнів може містити похибки, але ці похибки мають бути контрольовані таким чином, щоб отримати правильний підсумковий результат пословного розпізнавання та/або смислової інтерпретації мовлення.

Такий підхід може бути використаний при створенні багатомовних систем автоматичного розпізнавання мовлення, у комбінації з підходом, що має на меті виключити залежність оброблення мовленнєвого сигналу від конкретної мови [2].

У попередніх дослідженнях на першому рівні розглядався узагальнений фонемний розпізнавач, який дає відповідь розпізнавання у вигляді N >> 1 кращих послідовностей фонем разом з їх акустичними складовими в умовах вільного порядку слідування фонем [1]. На другому рівні узагальнений пословний розпізнавач проводить постпроцесинг результатів фонемного розпізнавання попереднього рівня. При цьому були досягнуті прийнятні результати експериментальних досліджень на базі даних одного диктора в умовах розпізнавання окремо вимовлених слів.

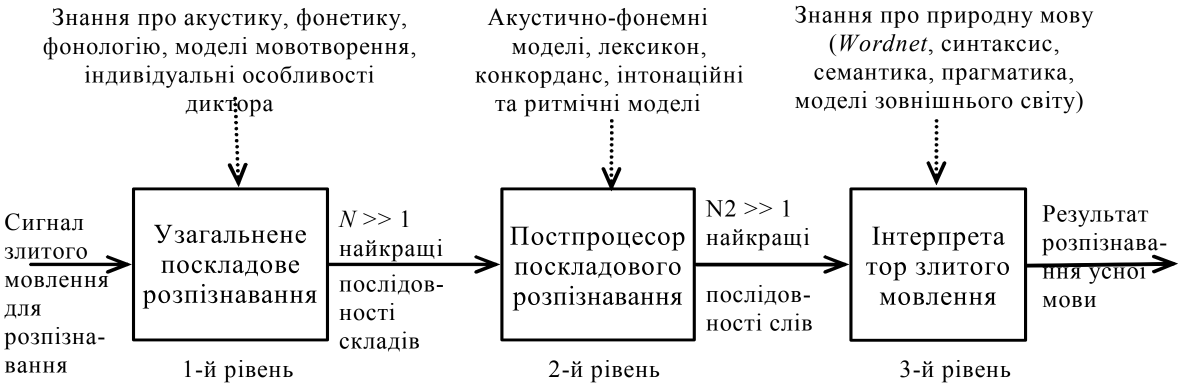


Рисунок 1 – Трирівнева структура системи поскладового розпізнавання усної мови

Надалі планувалося використовувати в експериментальних дослідженнях мовленнєву базу даних (корпус) кооперативу дикторів. Крім того, передбачалася інтеграція лексикону в процес обчислення значень змінних у вузлах графа постпроцесора з метою зменшення розгалужень на графі. Проблема вибору мовленнєвого образу, яким оперує постпроцесор, також залишалася відкритою.

У статті розглянуто поскладову модифікацію багатозначної трирівневої системи розуміння мовленнєвого сигналу. Структура цієї системи показана на рис. 1. Вона складається з трьох частин. Це узагальнений розпізнавач мовленнєвих образів на рівні нижчому, ніж слово, тобто складів, постпроцесор поскладового розпізнавання та інтерпретатор злитого мовлення.

Узагальнений поскладовий розпізнавач надає N >> 1 кращих відповідей розпізнавання в умовах вільної (або відносно вільної) поскладової граматики. Потім постпроцесор поскладового розпізнавання аналізує результати першого рівня, щоб належним чином згенерувати N2 >> 1 можливих послідовностей слів. Виходячи з цих послідовностей слів, інтерпретатор злитого мовлення приймає рішення стосовно смислу, який передається, з використанням знань про природну мову.

У розділі 1 ми обґрунтовуємо вибір мовленнєвих образів для першого рівня на основі складів. У розділі 2 описується постпроцесор, засобами якого здійснюється перехід від послідовностей складів до послідовностей слів. У розділі 3 ми характеризуємо базу даних і знань, яка використовується при розпізнаванні. Розділ 4 присвячено експериментальним дослідженням.

1. Вибір мовленнєвого образу для першого рівня

У попередніх роботах при побудові граматики на першому рівні обмежень на порядок слідування фонем не накладалося. Не дивлячись на швидке виконання, робасність системи розпізнавання є далекою від бажаної, особливо для кооперативу дикторів. Тепер пропонується розглядати склади як альтернативні мовленнєві образи, які все ще слабо залежать від словника.

Проаналізовані два шляхи вибору складів: на основі правил складоподілу та відкриті склади.

Вибір на основі правил складоподілу випливає з евристичних тверджень лінгвістичної науки щодо розміщення меж складів в залежності від сполучень фонем. Відкриті склади закінчуються голосним звуком або фонемою-паузою. Вибір складів на основі масиву даних також знаходиться у сфері інтересів і планується в подальших дослідженнях.

Словники складів були сформовані автоматично на базі частотного словника української мови обсягом у 137 640 слів. Хоча порядок слідування складів вільний, все ж на відкриті склади накладається додаткове обмеження: склади, які закінчуються фонемою-паузою, завжди слідують за складом, що закінчується голосним звуком.

Таблиця 1 – Фонемна коректність при розпізнаванні в залежності від типу мовленнєвого образу для різних україномовних навчальних вибірок

Навчальна вибірка	Тип мовленнєвого образу	Кількість моделей мовленнєвого образу	Фонемна коректність
11000 слів	монофон	55	46,0
11000 слів	склад на основі правил складоподілу	9 436	79,5
11000 слів	відкритий склад	4 966	78,3
100 речень	монофон	55	49,3
100 речень	склад на основі правил складоподілу	9 436	56,8
100 речень	відкритий склад	4 966	55,5

Табл. 1 ілюструє, що для окремо вимовлюваних слів поскладова граматика істотно покращує коректність пофонемного розпізнавання (до 1,6 разів) порівняно з розпізнаванням в умовах вільного порядку слідування фонем. Середня довжина українського слова складає 7,43 фонеми і максимальна – 20 фонем. В усіх випадках розпізнавання проведене з використанням системи Julius-Julian [3], виконується у реальному часі, який приблизно однаковий для розглянутих видів складів. Слід також зазначити, що склади, вибрані на основі правил, дають кращий результат.

2. Моделювання процедури постпроцесингу

На виході поскладового розпізнавача ми маємо $N >> 1$ кращих послідовностей складів, яким відповідає послідовність фонем $\Phi^r_{0Q^r} = \left(\varphi^r_1, \varphi^r_2, ..., \varphi^r_u, ..., \varphi^r_{Q^r}\right)$, $r = 1 : N$, де Q^r – довжина r-ї спостережуваної послідовності. Крім того, в результаті виконання першого рівня, кожна φ^r_u супроводжується оцінками акустичних параметрів, таких як тривалість фонеми d^r_u, її ймовірність ΔF^r_u і, можливо, іншими параметрами (енергія, рух основного тону тощо). Фактично ми розглядаємо послідовність фонетично-акустичних подій, які спостерігаються після застосування поскладового розпізнавання.

Метою постпроцесора є отримати для всіх $\Phi^r_{0Q^r}$, $r = 1 : N$ загалом $N1 >> 1$ прихованих послідовностей фонем $\Psi^{r1}_{0Q^{r1}} = \left(\psi^{r1}_1, \psi^{r1}_2, ..., \psi^{r1}_s, ..., \psi^{r1}_{Q^{r1}}\right)$, $r1 = 1 : N1$, $\psi \in \Psi \equiv \Phi$, і поставити їм у відповідність послідовності слів $J^{r2}_{0Q^{r2}} = \left(j^{r2}_1, j^{r2}_2, ..., j^{r2}_k, ..., j^{r2}_{Q^{r2}}\right)$, $r2 = 1 : N2$, $N2 >> 1$ і $j^{r2}_k \in J$, де J – лексикон. Щоб уникнути втрати фактичних послідовностей слів, зберігаємо $N2 >> 1$ відповідей розпізнавання.

Таким чином, ми інтерпретуємо спостережувані підпослідовності фонем $\Phi^r_{u_{s-1}u_s} = \left(\varphi^r_{u_{s-1}+1}, \varphi^r_{u_{s-1}+2}, ..., \varphi^r_{u_s}\right)$, $u_s - 1 \le u_s$, як перетворену приховану s – у фонему ψ^{r1}_{ks} регулярної транскрипції k-ї слова $j_{0q_k} = \left(\psi^{r1}_{k1}, \psi^{r1}_{k2}, ..., \psi^{r1}_{ks}, ..., \psi^{r1}_{kq_k}\right)$. Ймовірність того,

що спостережувана підпослідовність $\Phi^r_{s_{k-1}s_k} = \left(\varphi^r_{s_{k-1}+1}, \varphi^r_{s_{k-1}+2}, ..., \varphi^r_{s_k}\right)$, де $(s_k - s_k - 1) = l$ – довжина спостереження, є реалізацією прихованої транскрипції k-го слова $j_{0q_k} = \left(\psi^{r1}_{k1}, \psi^{r1}_{k2}, ..., \psi^{r1}_{ks}, ..., \psi^{r1}_{kq_k}\right)$, виражається добутком незалежних спотворень, що максимізується за границями $\{u_s\}$ прихованої фонеми ψ^{r1}_{ks}:

$$P\left(\Phi^r_{s_{k-1}s_k} \big/ j_{0q_k}\right) = \max_{\{u_s\}} \prod_{s=1}^{q_k} P\left(\Phi^r_{u_{s-1}u_s} \big/ \psi^{r1}_{ks}\right). \qquad (1)$$

Тут кожний множник $P(\Phi_{\mu\nu} / \psi)$ дорівнює 0, якщо $\Phi_{\mu\nu} = (\Phi_{\mu+1}, \Phi_{\mu+2}, ..., \Phi_\nu)$ не пов'язаний з прихованою ψ. В іншому випадку множник $P(\Phi_{\mu\nu} / \psi)$ обчислюється як функція $\Phi_{\mu\nu}$ і ψ, що враховує частотність та параметри нормального закону, яким описуються акустичні складові (d^r_u, ΔF^r_u тощо).

Кожна послідовність фонетико-акустичних подій обробляється запропонованим фільтром засобами динамічного програмування, як це показано на рис. 2. Тут ми спостерігаємо послідовність фонем (фонетичних подій) $(\varphi_1, \varphi_2, \varphi_3, \varphi_4)$, отриманих поскладовим розпізнавачем, за умови чотирьох прихованих фонем, представлених їх акустично-фонетичними моделями (АФМ). Прихована фонема може бути замінена однією, двома або трьома спостережуваними фонемами або бути пропущеною, тобто спостерігатися як порожня фонема $\varnothing$. Пунктирні лінії означають переміщення між моделями, що описуються граматикою, яка генерується на основі лексикону. Суцільні лінії показують внутрішні детерміновані переміщення.

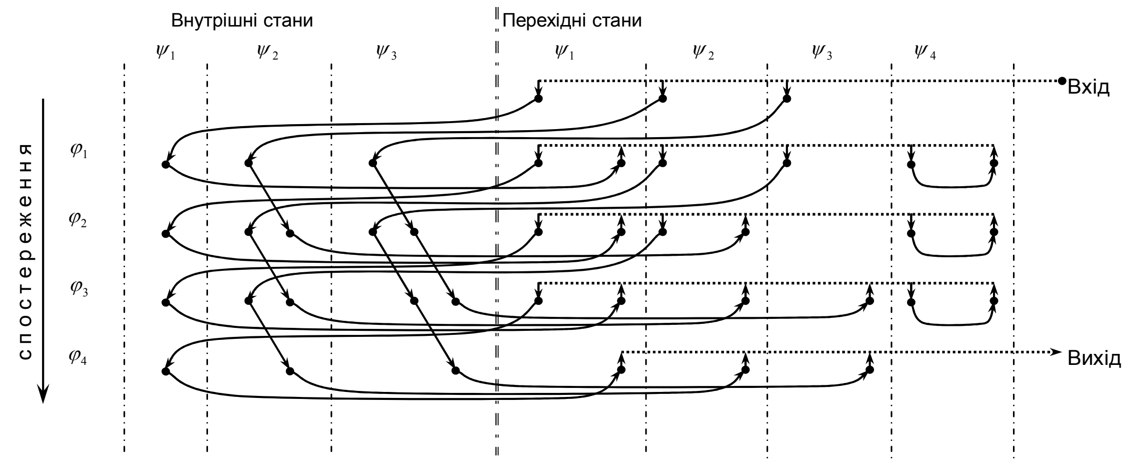


Рисунок 2 – Граф постпроцесингу для чотирьох фонетичних подій

Лексична складова постпроцесора забезпечує обмеження на послідовності фонем і виконання завершального перетворення фонем на графеми включно з визначенням границь слів.

Таким чином, $N >> 1$ кращі послідовності фонем, отримані в результаті дії першого рівня, перетворюються на $N2 >> 1$ послідовностей слів.

Параметри ймовірності (1), які також є параметрами АФМ, оцінюються за навчальною вибіркою згідно з [4]. На рис. 3 ми ілюструємо граф побудови потенційних АФМ для відповіді розпізнавання «pau k s1 o o pau» за умови вимовленого

слова «pau ts1 o h o1 pau» («цього»). Оптимальні траєкторії відображені суцільними лініями, а частина допустимої, але не оптимальної траєкторії позначена пунктирною лінією. Тут отримано такі фонемні описи для прототипів:

1:(PAU, k / pau, 4), 2:(PAU, k / pau, 5), 3:(PAU / pau, 6); 4:(k / 1, ts1, 7), 5:(k, s / 2, ts1, 8), 6:(s / 3, ts1, 9); 7:(s, O / 3, e1, 9), 8:(o / 4|5, O, 9); 9:(∅ / 6|7|8, h, 11); 10:(o / 6|7|8, h, 12); 11:(o / 9, o1, 13); 12:(∅ / 10, o1, 13); 13:(PAU / 11|12, pau).

У дужках перед похилою рискою позначена послідовність фонем, яка замінює модельну фонему. Кожна фонема з послідовності асоціюється зі станом моделі, а фонема, позначена великою літерою, асоціюється з тим станом, який позначає фонеми, що збігаються зі спостережуваними. Праворуч від похилої риски зазначено ім’я моделі фонеми і ідентифікатори сусідніх модельних прототипів. Додатково зазначається ймовірність кожного модельного прототипу.

Зауважимо, що акустичні складові моделей поновлюються для кожного прототипу з метою побудови глобальної моделі ітераційним шляхом або шляхом вилучення менш імовірних моделей.

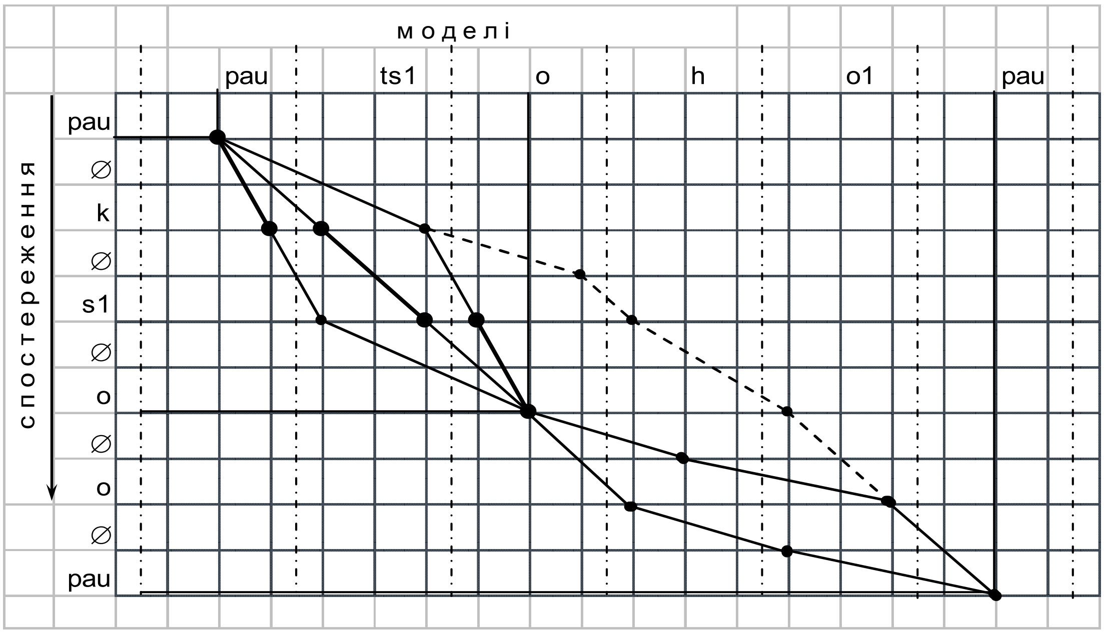


Рисунок 3 – Граф побудови потенційних прототипів акустико-фонетичних моделей за навчальною вибіркою

3. База даних і знань

База даних і знань включає україномовний мовленнєвий корпус для оцінки параметрів акустичних і акустично-фонетичних моделей (АФМ) для формування лексикону.

Ми використовували україномовний багатодикторний мовленнєвий корпус, який містить понад 30 000 реалізацій слів і тисячі речень близько 100 дикторів, що мешкають у різних областях України. Реалізації зберігають частотні пропорції фонем і є фонетично збалансованими [5].

Лексикон містить близько 2 мільйонів словоформ, яким відповідають 151 000 основних форм (лем). Фактично цим лемам відповідає більше 3 мільйонів словоформ, але у багатьох з них однакова орфографія і вимова [5].

На основі лексикону та текстового корпусу обсягом 250 МБ було згенеровано частотний словник на 157 000 слів.

Модуль перетворення фонемного тексту на орфографічний використовує 22 узагальнених правила відображення символьних послідовностей і звертається до всього лексикону. Фактично на стадії обчислення вузлів на графі (рис. 1) замість цього модуля можна використати асоціативні масиви послідовностей фонем (2- і 3-грами), що трапляються в середині слів.

4. Експериментальні дослідження

Експеримент ділився на стадії (1) підготовка навчальної та контрольної вибірок, (2) фонемне навчання, (3) поскладове розпізнавання, (4) оцінка параметрів постпроцесора, (5) тестування постпроцесора.

Для навчання фонем вибірка формувалась з реалізацій україномовного багатодикторного мовленнєвого корпусу. Ми розглядали окремі (ізольовані) слова для оцінки параметрів акустичних моделей фонем. Загалом ми мали близько 19 858 реалізацій слів і 147 445 реалізацій фонем, окрім фонеми-паузи, від 70 дикторів.

Алфавіт, який використовувався, містить 55 базових фонем (монофонів). Серед них як наголошені, так і ненаголошені варіанти голосних звуків, варіанти, що палаталізують (пом'якшують) приголосні звуки і фонема-пауза. Частотність кожної фонеми не-паузи у навчальній вибірці знаходиться між 30 (пом'якшені 'sh' і 'zh') і 1200 для ненаголошеної 'o'. Модель короткої паузи не була передбачена, оскільки навчальна вибірка включає лише ізольовані слова.

Акустичні моделі були навчені і вдосконалені з використанням програмного комплексу HTK [6] для кожного з 55 монофонів і фонеми-паузи. При вдосконаленні були враховані акустична мінливість і частотність фонем. Отримані моделі фонем мають кожна три стани і від 4 до 12 гауссівських сумішей.

Поскладове розпізнавання було здійснене з використанням програмного комплексу Julian [3] для двох наборів складів: відкриті склади і склади на основі правил з відповідними граматиками згідно з розділом 2.

Відповідь поскладового розпізнавання, включно з сегментацією і критерієм фонем, використовувалася для оцінки параметрів акустично-фонетичних моделей (АФМ). Для цього ми скористалися усномовними даними одного диктора, які не використовувалися при побудові акустичних моделей фонем. Ці усномовні дані складалися зі слів, отриманих в результаті сканування частотного словника, починаючи з найбільш частотного слова, і відбору тих слів, що містять нові трифони. Загалом таких слів було записано 8 000, серед яких перших 3 000 записано двічі. Ми провели оцінку різної кількості моделей для різних початкових умов, як описано в табл. 2.

Таблиця 2 – Результати процедури постпроцесингу

Тип складу	АФМ навчального корпусу	Загальних/ використаних АФМ	Тестовий корпус	N2	Словесна похибка, %
На основі правил	5 000	3 700 / 3 300	6 000 слів	5	4,7
Відкритий	5 000	3 900 / 3 300	6 000	5	5,2
На основі правил	5 000	3 700 / 3 300	6 000	7	4,5
Відкритий	5 000	3 900 / 3 300	6 000	7	4,8
На основі правил	11 000	7 500 / 7 000	2 100*3	5	4,9
Відкритий	11 000	7 900 / 3 300	2 100*3	5	5,1
На основі правил	11 000	3 700 / 3 300	100 речень	5	18,2

Перед перевіркою постпроцесора ми скорегували його лексичні параметри: обмеження щодо наголошеної голосної були зняті і у відповіді розпізнавання допускалася послідовність слів. У всіх експериментах використано повний словник (2 млн словоформ).

Постпроцесор був випробуваний на різних наборах ізольованих слів, а відповідь розпізнавання отримувалась у формі N2-кращих послідовностей слів.

З усього випливає, що точність постпроцесора складає близько 95 % для ізольованих слів. Істотне погіршення на реченнях викликано численними короткими словами з високим критерієм. Штраф на переміщення між словами на графі постпроцесора, можливо, поліпшив би результат.

Висновки

Запропонована модель є актуальною для мов з великою кількістю словоформ та відносно вільним порядком слідування слів, до яких відносяться і слов'янські мови.

Більш відповідна акустична модель для розпізнавання мовлення – це фонемно-трифонна модель, оскільки враховується явище коартикуляції. Фонемно-трифонна модель оперує $|\Phi|3$ породжувальними граматиками, і обсяг обчислень зростає у $|\Phi|2$ разів, порівняно з монофонною моделлю. Окрім цього, оброблення фонемно-трифонної граматики потребує значно більше обмежень, що призводить до додаткових обчислень. Таким чином, доцільно вибрати значення N, близьке до числа $|\Phi|$, і навіть більше з метою досягнення оптимальних витрат пам'яті і обчислення.

Проблема залишається в тому, яким чином запобігти втраті оптимального розв'язку.

Недоліком кожного постпроцесора є його активація після закінчення базового процесу. Отже, слід розглянути шляхи інтеграції постпроцесора в обчислення вузлів графа поскладового розпізнавання.

Автоматичне формування множини складів або альтернативних мовленнєвих образів нижчого, ніж слово, рівня за масивом даних є метою подальших досліджень.

Література

1. Taras K. Vintsiuk, Mykola M. Sazhok. Multi-Level Multi-Decision Models in ASR // Proc. of the 10th International Workshop «Speech and Computer», SPECOM'2005. – Patras. – 2005. – P. 69-76.
2. Gérard Chollet, Kevin McTait, Dijana Petrovska-Delacrétaz. Data Driven Approaches to Speech and Language Processing // G. Chollet et al. (Eds.): Nonlinear Speech Modeling, LNAI 3445. – 2005. – P. 164-198.
3. Lee, T. Kawahara and K. Shikano. Julius – an open source real-time large vocabulary recognition engine // In Proc. European Conference on Speech Communication and Technology (EUROSPEECH). – 2001. – P. 1691-1694.
4. Mykola Sazhok. Generative Model for Decoding a Phoneme Recognizer Output // Proc. of the 8th International Conference «Text, Speech and Dialogue», TSD'2005. – Karlovy Vary. – 2005. – P. 288-293.
5. Nina Vasylyeva, Mykola Sazhok. Text Selection for Training Procedures under Phoneme Units Variety // Proc. of the 10th International Workshop «Speech and Computer», SPECOM'2005. – Patras. – 2005. – P. 629-632.
6. Широков В., Монако В. Організація національної лексикографічної мережі // Мовознавство. – № 5. – 2001.
7. Young S.J. et al., HTK Book, version 3.1. – Cambridge University, 2002. – 355 p.

Н.Б. Васильева, Н.М. Сажок

Моделирование многоуровневого послогового распознавания речевого сигнала

В статье проводится распространение многоуровневой многозначной модели автоматического распознавания слитной речи на случай послогового распознавания. Рассматриваются два уровня из трех. На первом уровне проводится распознавание в условиях послоговой грамматики, на втором уровне проводится обработка (постпроцессинг) исходных данных первого уровня с целью получения соответствующих последовательностей слов. В описанной модели постпроцессинга обращается внимание на полученные оценки акустических составляющих речевого сигнала, а последовательность и фонетические особенности вместе с лексиконом. Анализируются пути выбора множества единиц на слоговом уровне речевых образов. Описывается многодикторный речевой корпус и лексикон, которые использованы в экспериментальном исследовании. Обсуждаются результаты экспериментов, проблемы и будущие исследования.

Стаття надійшла до редакції 23.07.2008.